

# Database Research Self Assessment Lowell, MA (2003)

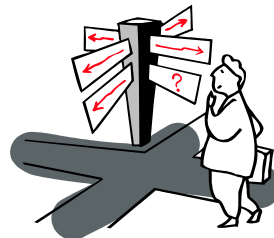
15.1.2004

Felix Naumann

Jens Bleiholder

## Outline

- The meeting (where, when)
- The participants (who)
- The topics (what)



## Purpose of the meeting

- Assessment/survey of recent and current DB research
  - Objective (as far as possible)
  - Self-critical
- Recommend problems and areas that deserve additional focus
  - No „grand challenge“, DB is part of all „grand challenges“
- Done by „senior researchers“
  - Experienced
  - Know the field

## Previous meetings

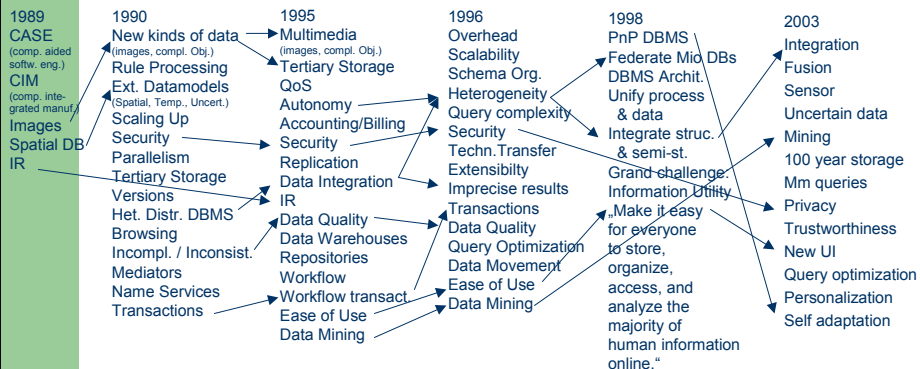
- Laguna Beach, 1989
- Palo Alto, 1990 („Lagunita“)
- Palo Alto, 1995
- Cambridge, 1996
- Asilomar, 1998
- Reports published in:
  - SIGMOD Record, CACM, ACM CSUR, WWW, ...

# Program

- Gong-Show (Su evening)
- Talks, different topics
- Breaks, discussions
- Write report
- 2 days (Mo/Tu)



# Topics over time



## The participants

- Serge Abiteboul
- Martin Kersten
- Rakesh Agrawal
- Michael Pazzani
- Phil Bernstein
- Mike Lesk
- Mike Carey
- David Maier
- Stefano Ceri
- Jeff Naughton
- Bruce Croft
- Hans Schek
- David DeWitt
- Timos Sellis
- Mike Franklin
- Avi Silberschatz
- Hector Garcia Molina
- Rick Snodgrass
- Dieter Gawlick
- Mike Stonebraker
- Jim Gray
- Jeff Ullman
- Laura Haas
- Gerhard Weikum
- Alon Halevy
- Jennifer Widom
- Joe Hellerstein
- Stan Zadonik
- Yannis Ioannidis

7



Forschungsseminar, WS 03/04

15.1.2004

## Serge Abiteboul

- INRIA
- XML, XML, XML
- Scientific Advisor: Xyleme
- Bücher: Data on the Web, Foundations of Databases (Theorie-Klassiker)
- <http://www-rocq.inria.fr/~abitebou/>
- DBLP index:



8



Forschungsseminar, WS 03/04

15.1.2004

## Martin Kersten

- CWI: Centrum voor Wiskunde en Informatica, Nederlande
- DBMS, XML
- Abteilungsleiter mit > 50 Mitarbeitern
- <http://homepages.cwi.nl/~mk/>
- DBLP index:



## Rakesh Agrawal

- IBM Fellow
- IBM Almaden
- „Inventor“ of Data Mining
- Now: Privacy of IS
- <http://www.almaden.ibm.com/u/ragrawal/bio.html>
- DBLP index: 132



## Michael Pazzani

- UC Irvine
- AI / Machine Learning
- <http://www.ics.uci.edu/~pazzani/>
- DBLP index: 90



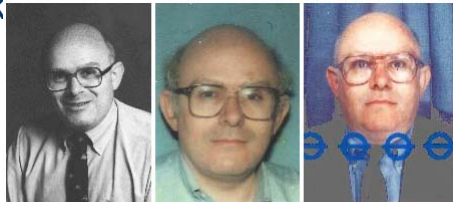
## Phil Bernstein

- Microsoft Research, Redmond
- Affiliate Professor @ UW
- ACM Fellow
- Architect for Matadata Services for MS SQL Server
- Now: Model Management
- <http://www.research.microsoft.com/~philbe/>
- DBLP index: 107



## Mike Lesk

- Rutgers University
- ACM Fellow
- Digital Library, IR
- <http://lesk.com/mlesk/>
- DBLP index:34



Professional

Amateur

Coin-operated

13

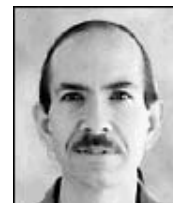


Forschungsseminar, WS 03/04

15.1.2004

## Mike Carey

- BEA Systems
- Zuvor: U Wisconsin, IBM Almaden, Propel
- DBLP index: 146



14



Forschungsseminar, WS 03/04

15.1.2004

## David Maier

- Early work: Relational Theory
- OGI, Oregon
- <http://www.cse.ogi.edu/DISC/people/maier.html>
- DBLP index: 148



## Stefano Ceri

- Distributed DBMS, Active DB
- [Politecnico di Milano](http://www.elet.polimi.it)
- <http://www.elet.polimi.it/upload/ceri/>
- DBLP index: 160



## Jeff Naughton

- NIAGARA project
  - XML DBMS
- U Wisconsin-Madison
- <http://www.cs.wisc.edu/~naughton/naughton.html>
- DBLP index: 109



## Bruce Croft

- IR and Digital Libraries
- University of Massachusetts – Amherst
- ACM Fellow
- <http://ciir.cs.umass.edu/personnel/croft.html>
- DBLP index: 114



## Hans-Jörg Schek

- Hyperdatabases
  - DB functionality at high abstraction level (services etc.)
- Swiss Federal Institute of Technology Zurich
- ETH Zentrum
- DBLP index: 139



## David DeWitt

- DBMS, Optimierung
- Niagara
- U Wisconsin
- ACM Fellow
- <http://www.cs.wisc.edu/~dewitt/>
- DBLP index: 141



## Timos Sellis

- National Technical University of Athens
- <http://www.dbnet.ece.ntua.gr/~timos/index.html>
- DBLP index: 100



## Mike Franklin

- Distributed DB, Mobile/Pervasive Computing, Data Streams
- UC Berkeley
- DBLP index: 91



## Avi Silberschatz

- Networks, Multidatabases, Privacy, ...
- Formerly Bell Labs, Now Yale
- ACM and IEEE Fellow
- <http://www.bell-labs.com/user/avi/>
- DBLP index: 179



## Hector Garcia Molina

- TSIMMIS, P2P
- Stanford
- <http://www-db.stanford.edu/people/hector.html>
- ACM Fellow
- DBLP index: 268



## Richard Snodgrass

- Temporal DBMS
- Sehr aktiv in der ACM, ACM Fellow
- DBLP index: 162



## Dieter Gawlick

- Oracle
- DBLP index: 12

## Mike Stonebraker

- Ingres, POSTGRES, Illustra
- UC Berkeley
- [http://epoch.cs.berkeley.edu:8000/nasa\\_e2e/mike.html](http://epoch.cs.berkeley.edu:8000/nasa_e2e/mike.html)
- DBLP index: 196



27

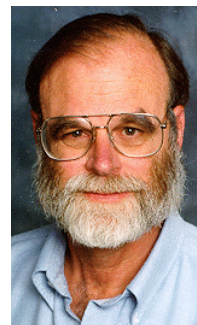


Forschungsseminar, WS 03/04

15.1.2004

## Jim Gray

- Microsoft Research, Bay Area (BARC)
- Transactions, Astronomy
- ACM Fellow, Turing Award 98
- DBLP index: 89



28



Forschungsseminar, WS 03/04

15.1.2004

## Jeff Ullman

- Stanford University
- Früher: Algorithmen und Theorie
- Buch Klassiker: Principles of Database and Knowledge-Base Systems
- <http://www-db.stanford.edu/~ullman/>
- DBLP index: 231



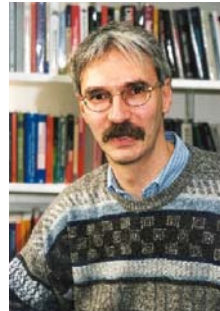
## Laura Haas

- IBM Almaden & SVL
- Clio, Garlic, R\* Distributed DBMS
- <http://www.almaden.ibm.com/cs/people/laura/>
- DBLP index: 52



## Gerhard Weikum

- MPI für Informatik, Saarbrücken
- QoS, autonomic computing, distributed DBMS, workflow & web services
- <http://www.mpi-sb.mpg.de/~weikum/>
- DBLP index: 148



31



Forschungsseminar, WS 03/04

15.1.2004

## Alon Halevy

- U Washington, Seattle
- TukWila, Information Manifold, Piazza (P2P), Strudel (Web-site management)
- Query Answering Using Views
- DBLP index: 34 seit 2000, 77 zuvor



32



Forschungsseminar, WS 03/04

15.1.2004

## Jennifer Widom

- Stanford
- Data Streams, Lore
- <http://www-db.stanford.edu/~widom/>
- DBLP index:123



## Joe Hellerstein

- UC Berkeley
- P2P, TinyDB (Sensors), ABC Potters Wheel
- <http://db.cs.berkeley.edu/~jmh/>
- DBLP index: 66



## Stan Zdonik

- Brown University (RI)
- <http://www.cs.brown.edu/people/sbz/>
- DBLP index: 86

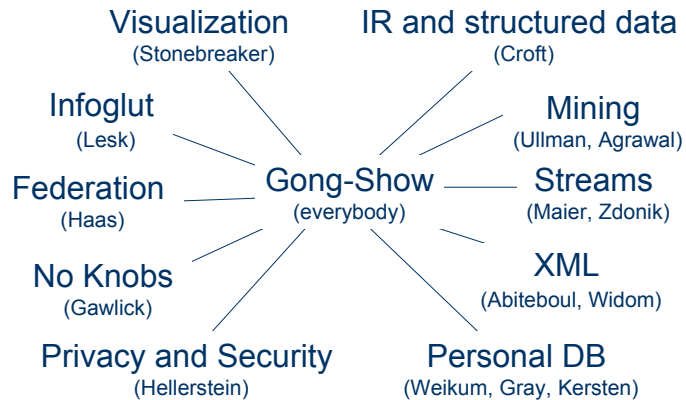


## Yannis Ioannidis

- University of Athens
- Query Optimization / Scheduling, Histograms (VLDB Award)
- DBLP index: 86



## Talks



37



Forschungsseminar, WS 03/04

15.1.2004

## Summary of topics

- Integration of data types
- Information Fusion
- Sensor data/networks
- Uncertain data
- Data Mining
- 100 year storage
- MM queries
- Personalization
- Self adaptation
- Privacy
- Trustworthiness
- New UI
- Query optimization

38



Forschungsseminar, WS 03/04

15.1.2004

## Integration of new data types

- New data types due to WWW
  - Text-, image-, sound data
  - Geographical and temporal data
  - Code and Streams
- Extensions of existing DBMS
  - E.g. OO, OR, user defined procedures (code)
  - But: Minimal, „second-class“

## Integration of new data types

- Totally new DBMS core (rethink DBMS)
- More and new „first class“ objects
  - Structured data
  - Text-, image-, sound data, geographical, temporal data
  - Procedures (data types and methods)
  - Triggers, streams
  - Suitable query language
- Research vs. industry
- 5 year goal: first prototypes

## Information Fusion

- Integration/fusion/merging of data
- Federation has been topic since 1989
- Within enterprises
  - ETL, feasible, manageable (10-100)
- Among enterprises
  - WWW, ad-hoc, data stays in sources/enterprises
  - on-the-fly, (up to 1 Millions of sources)

## Information Fusion

- Main problem: semantic heterogeneity
  - Mark Twain vs. Samuel Clemens
  - Euro vs. Dollar
- Probabilistic query answering
- Monitoring applications
  - „Let me know when one of my mileage plans is giving bonuses on hotel stays for hotels that have hotels near the sites of conferences or meetings I will be attending...“

## Sensor data and sensor networks

- Networks of many, many sensors
  - Measuring: temperature, location, ...
  - Low cost, self-powered, wireless
- New requirements
  - Use more power communicating than computing,  
⇒ do query processing in the sensors
  - The network is the *database*

## Sensor data and sensor networks

- Query plan changes with network changes
  - Sensors die or disconnect
  - Sensors regroup
- Information integration
  - Calibration, uncertain data
  - Deduce high level facts from low level data, e.g.  
locate a person from heat, sound, vibration sensors

## Reasoning about uncertain data

- Traditional DBMS
  - Strings, numbers, business  $\Rightarrow$  exact
  - No need/facilities for approximate, imprecise data
- Other domains: imprecise
  - Especially scientific data, error is measurable
  - Location data, similarity measures
  - Lineage/provenance needed
- Need for stochastic query processing
  - Stochastic answer, evidence accumulation
  - Imprecise query

## Data Mining

- Discover models of data sets
  - Classification, clustering, association rule discovery, summarization
  - Part of mainstream DBMS
- Only one question
  - „Tell me something interesting“
- Challenges
  - Background process in DBMS
  - Integrated in systems, couple with triggers
  - Teach people how to properly use these tools

## 100 year storage

- Information digitally stored
- Information disappears
  - Deteriorating media
  - Obsolete device to read data
  - Old data format, old application
- Need for permanent migration
- Need for storing metadata
  - DM, EUR oder BRD/DDR

## 100 year storage

- Personal information
  - Emails, videos, appointments
  - PDA, office, at home
  - Keep them a lifetime
- One information system for all
  - Automatic metadata management
  - Automatic migration
  - Access all data, anytime

## Other topics

- Multimedia queries
  - Increasing amount of multimedia data, especially personal data
  - How do I access the data?
- Personalization
  - Answer dependent of role
  - Relevance feedback dependent of role
  - Offer information for different purposes
  - What are appropriate metadata?

## Other topics

- Privacy
  - Much data available, correlations possible
  - Access by user AND usage
  - Include purpose description in query
- Trustworthy System
  - Protect data from loss, disclosure
  - Grant always access to authorized users
  - Digital rights management
  - Correctness of answers

## Other topics

- Self adaptation
  - Less competent DB administrators
  - Simplify: „no more knobs“
  - Detect malfunctions, recover
- New user interfaces
  - From SQL to Xquery, and now? Ontologies?
  - Visualization
- Query optimization
  - Integration, semistructured, stream processors, sensor networks

## One meeting, one opinion?

- „So what's the message? Laura says it's impossible and Stonebreaker says it's done.“
- Information integration – at what level?
- Do „web services“ help at solving the „semantic heterogeneity“ problem?

## Next steps

- Design a testbed for information integration
  - Like TREC in IR
  - How to do that?
- Connect to new areas, be open
- Next meeting with younger researchers

## Information sources

- Report
  - WWW: <http://research.microsoft.com/~Gray/Lowell/>
  - Panel SIGMOD 2003: Report on the Fourth "Where Should We Be Going" Meeting (Asilomar++)
- Transcript
- There are also reports from previous meetings
- Questions?

## Bibliography

1. Laguna Beach, CA, 1989
  - SIGMOD Record 18(1), 17-26 (1989)
2. Palo Alto, CA ("Lagunita"), 1990
  - <http://doi.acm.org/10.1145/125223.125272>
3. Palo Alto, CA, 1995
  - <http://doi.acm.org/10.1145/381854.381886>
4. Cambridge, MA, 1996
  - <http://doi.acm.org/10.1145/242223.242295>
5. Asilomar, CA, 1998
  - <http://doi.acm.org/10.1145/306101.306137>
6. Lowell, MA, 2003
  - <http://research.microsoft.com/~Gray/Lowell/>

## The end

- Questions?

## Talks

- Gong-Show (everybody)
- IR and structured data (Croft)
- Infoglut (Lesk)
- XML (Abiteboul, Widom)
- Federation (Haas)
- Mining (Ullman, Agrawal)
- Streams (Maier, Zdonik)
- Visualization (Stonebreaker)
- Personal DB (Weikum, Gray, Kersten)
- No Knobs (Gawlick)